# AN EVALUATION OF AN INDEX OF AFFINITY FOR COMPARING ASSEMBLAGES, IN PARTICULAR OF FORAMINIFERA

*by* M. J. ROGERS

ABSTRACT. An index of affinity which is related to the city block metric is described. Its values are found to be comparable with values derived from Pearson's coefficient of contingency when model sample assemblages are used. The value of the index of affinity which indicates that samples are taken from the same assemblage varies with the type of assemblage being sampled. This critical value can be calculated by using a simple formula. Analyses of samples of recent foraminiferal assemblages from the coast of California and the Western Approaches to the English Channel agree, in general, with results obtained using traditional techniques and factor analysis. The index of affinity is very simple to calculate and is equally effective if one or many pairs of samples are being compared.

A NATURAL assemblage is not always a homogeneous unit and the species are not always randomly distributed. As Buzas (1965) found, the chi-square test showed that two samples, taken from the same place at the same time, apparently were samples from different assemblages. This is due to the test's sensitivity to lack of homogeneity between samples. A simple similarity measure, here called the index of affinity, has been used by Sanders (1960) in work on the soft-bottom community and Murray (1973) in work on Foraminifera, but the meaning of its values has been only intuitively understood as yet. The index of affinity is derived from the Manhattan or city block metric which has been used in palynology by Gordon and Birks (1972, 1974) and Birks (1974), and is closely related to the mean character difference used by Cain and Harrison (1958) and Czekanowski (1909). It is the purpose of the present paper to show how the index of affinity varies under certain circumstances and to see whether analyses based on it are approximately as effective as those based on more rigorous analyses such as chi-square and factor analysis.

The variation of the index of affinity was investigated using artificial assemblages. Samples were taken from these assemblages using a random-number generator. The index of affinity and a coefficient based on chi-square were calculated for each sample pair from a particular assemblage. The frequency distribution for the two measures of similarity was then found for each artificial assemblage. The results obtained using the artificial assemblages were then applied to a series of samples taken from two very different natural assemblages. A comparison was made between factor analysis (Streeter 1972) and an analysis based on the index of affinity using Phleger's (1964) Gulf of California samples.

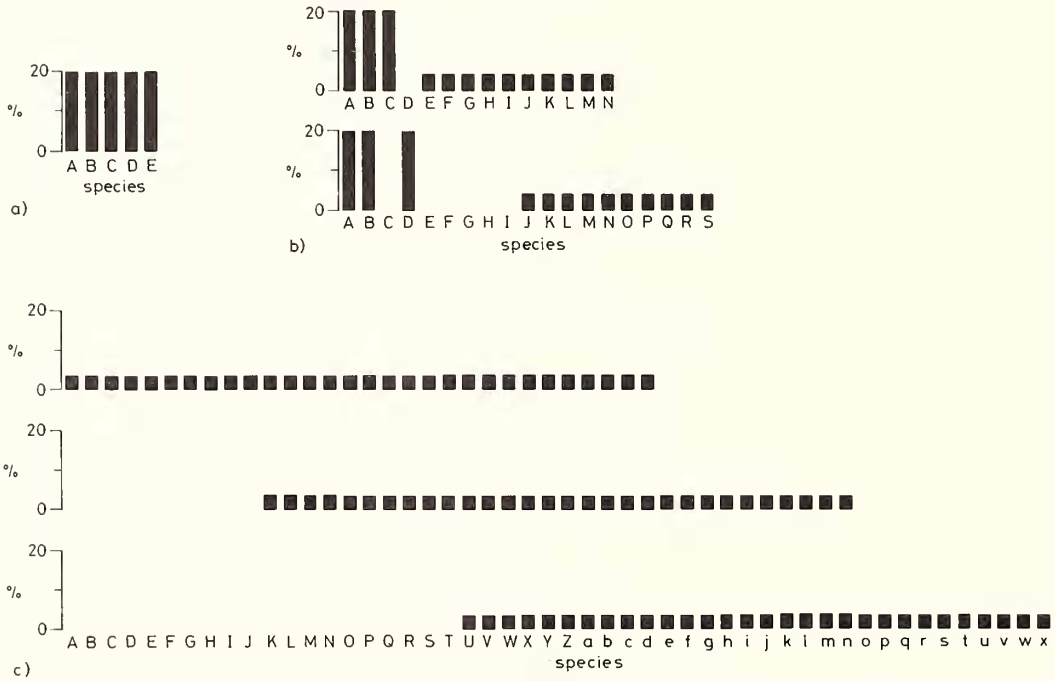
Once a measure of similarity was accepted, there still remained the problem of grouping sample assemblages. This is a trivial problem when few samples are involved. For large numbers of samples cluster analysis may be used, and there are various methods of grouping or splitting collections of samples. For useful summaries see Sokal and Sneath (1963), Sneath and Sokal (1973), and Cormack (1971). In this paper a pair-group method was used to see whether the sample assemblages

[Palaeontology, Vol. 19, Part 3, 1976, pp. 503–515.]

group readily. Inter-group and within-group frequency distributions were then drawn for the major groups. Using these methods it is possible to make certain statements about the assemblage (or assemblages) from which the samples were drawn and suggest areas where closer inspection of the data should be concentrated.

## METHOD

To test the distribution of the index of affinity and the coefficient of contingency 100 samples were taken from an artificial assemblage and each sample contained 250 random numbers. Five artificial assemblages were designed. The first assemblage contained five equally abundant species (text-fig. 1*a*). The range of random numbers was divided into five equal parts and as each of the 250 random numbers in a sample was generated it was assigned to the part (or species) which contained that number. In three other artificial assemblages, 10, 20, and 30 species were equally represented. The fifth artificial assemblage had three species of 20% abundance each and ten species of 4% abundance. These artificial assemblages could represent biocoenoses or thanatocoenoses from which random samples were drawn. The five-species assemblage might represent a restricted environment such as a marsh or desert.




TEXT-FIG. 1. Bar diagrams showing the distribution of species in three models. *a*, one artificial assemblage containing five equally abundant species. *b*, two artificial assemblages are shown, each containing three species at 20% abundance and ten species at 4% abundance. Two species at 20% abundance and five species at 4% abundance are common to both assemblages. *c*, three artificial assemblages shown out of a possible twenty. Each assemblage contains thirty equally abundant and adjacent species selected from a line of fifty available species.

The thirty-species assemblage could represent a more equable environment such as a shelf sea or temperate grassland. In fact, most natural environments have a few dominant species with varying numbers of less-abundant species and the last artificial assemblage may most nearly resemble a biocoenose.

Having used samples from the same artificial assemblage to establish how the value of the index of affinity and the coefficient of contingency varied with the composition of the assemblage, samples from different artificial assemblages were compared. For example, fifty samples were taken from an assemblage of five species and fifty samples were taken from an assemblage of ten species (five species being common to both assemblages). In the last pair of artificial assemblages two of the species of 20% abundance were common to all samples, as were five of the species of 4% abundance, but the third species of 20% abundance and five species at 4% abundance differed in the two blocks of fifty samples (text-fig. 1*b*). A final model was designed to represent a gradual change in the biocoenose from, for example, a mud community to a sand community. A total of fifty species (*A*-*x*) were present, but each artificial assemblage contained only thirty equally abundant and adjacent species. Text-fig. 1*c* shows three of the artificial assemblages. The two upper assemblages have twenty species in common as do the two lower assemblages, but the upper and lower (which are the most dissimilar) have only ten species in common.

The city block metric is defined as follows

$$\text{C.B.}_{kj} = \sum_{i=1}^{n} \left| X_{ki} - X_{ji} \right|$$

where $i = i$th species; $k, j =$ samples; $n =$ total number of species.

The index of affinity is calculated only on percentage data and, using the same notation, may be defined

$$\text{I.A.}_{kj} = \sum_{i=1}^{n} \min (X_{ki}, X_{ji}).$$

If the data are in percentage then C.B. varies from 0 to 200, and the index of affinity, which varies from 0 to 100, equals $(100 - \frac{1}{2}\text{C.B.})$.

The value of chi-square for two samples, $j$ and $k$, is given by the following equation

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_{ji} - EX_{ji})^2}{EX_{ji}} + \sum_{i=1}^{n} \frac{(X_{ki} - EX_{ki})^2}{EX_{ki}}$$

where degrees of freedom $= n - 1$; $X_{ji} =$ number of observations of $i$th species in sample $j$; $EX_{ji} =$ expected value of $X$ where

$$EX_{ji} = \frac{\sum_{i=1}^{n} X_{ji} \times (X_{ji} + X_{ki})}{\sum_{i=1}^{n} X_{ji} + \sum_{i=1}^{n} X_{ki}}.$$

To calculate the index of affinity directly from percentage data the lowest value for each species is summed over all species in the two samples, whereas to calculate

the city block metric the difference between the two values for a given species is summed.

i.e.

| species | sample $j$ | sample $k$ | % in common I.A. | % difference C.B. |
|---|---|---|---|---|
| A | 20 | 15 | 15 | 5 |
| B | 8 | 17 | 8 | 9 |
| C | 43 | 19 | 19 | 24 |
| D | 24 | 31 | 24 | 7 |
| E | 5 | 18 | 5 | 13 |
| Total | 100 | 100 | 71 | 58 |

$$\text{I.A.} = (100 - \tfrac{1}{2}\text{C.B.}) = (100 - \tfrac{1}{2} \times 58) \times 71$$

In order to compare the values of the index of affinity with values of chi-square it was necessary to express chi-square within the range 0–100. This was done using an extension of Pearson's coefficient of contingency as follows:

$$C = \left(\frac{\chi^2}{N+\chi^2}\right)^{\frac{1}{2}}.$$

$C$ = coefficient of contingency; $N = n_1 + n_2$, where $n_1$ = number of identifications in sample 1 and $n_2$ = number of identifications in sample 2.

Since $0 \leqslant \chi^2 \leqslant N$ (Kendall and Stuart 1967, p. 557) and the index of affinity ranges from 0 to 100, so $0 \leqslant (200 \times C^2) \leqslant 100$ will give a comparable range for an expression linked to $C^2$. Using the above definitions, a closely similar sample pair will have a high index of affinity and a low coefficient of contingency.

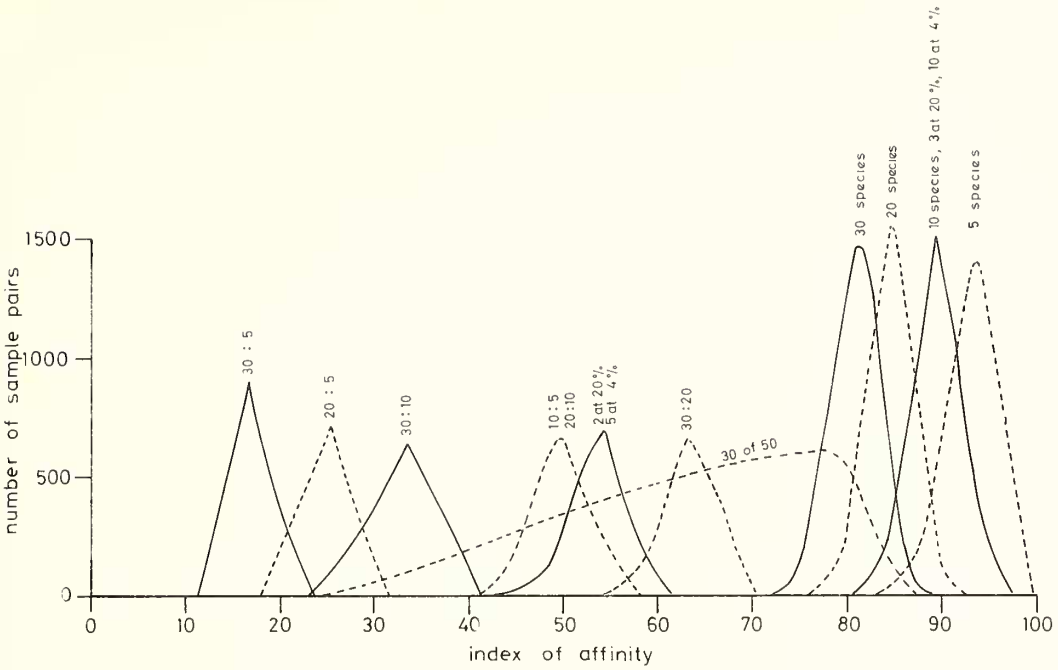
Thus $D = 100\text{–}200C^2$, ($D$ = modified coefficient of contingency) will give values of $D$ which lie within the same range as the values of the index of affinity, high values indicating close similarity of samples. The equations may be rewritten

$$\chi^2 = \frac{ND}{200 - D}$$

whereupon $ND/200 - D$ is distributed approximately as $\chi^2_{(k-1)}$ where $k$ = number of classes, and thus probability statements may be made using this statistic.

For each group of 100 samples, each sample was compared with every other sample and the similarity index entered in the lower diagonal of a square matrix. The distribution of the values was then recorded. When two types of artificial assemblages were compared, only the distribution of the indices for mixed pairs was considered (i.e. only 2500 pairs, 4950 being found for the 100 samples from one assemblage only). Text-fig. 2 shows the distributions of the index of affinity. Text-fig. 3 shows the distribution of the modified coefficient of contingency.

Cluster analysis was not applied to samples taken from the artificial assemblages but it was performed on samples from natural assemblages. The method used was as follows. The two samples which gave the highest index of affinity in the affinity matrix were placed in the first group. An average sample was formed for this group

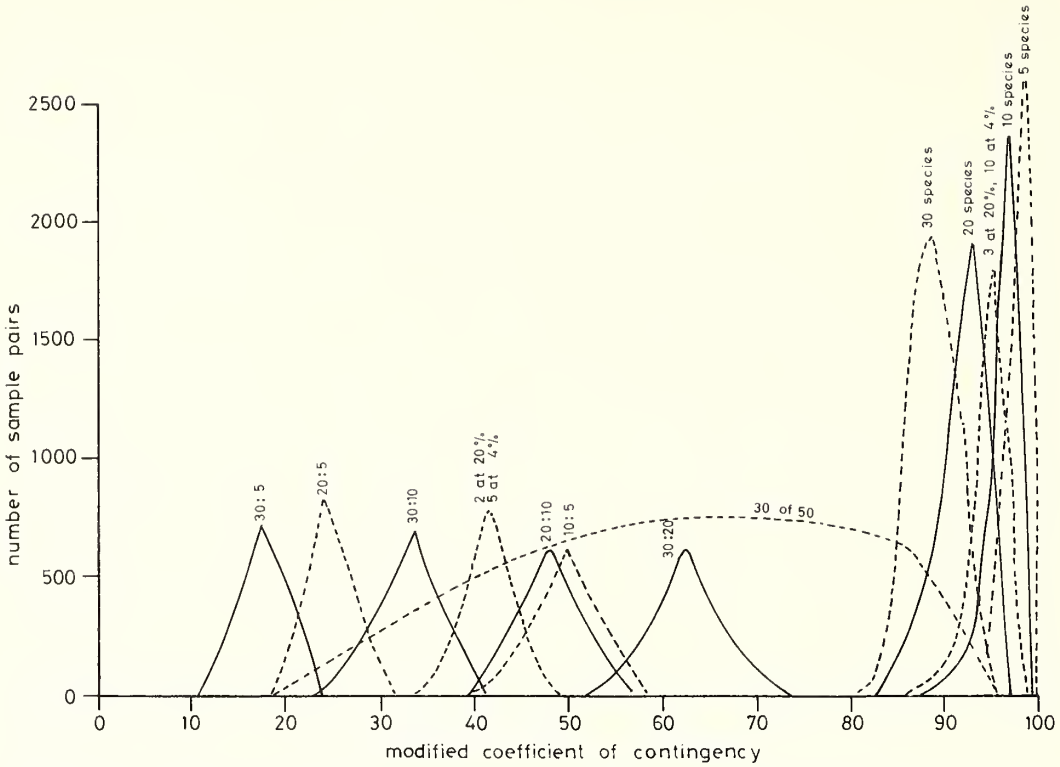



TEXT-FIG. 2. Curves showing the frequency distribution of the index of affinity for sample pairs in artificial assemblages. The numbers at the peak of each curve indicate the number of species in the assemblage(s). Unless otherwise indicated, the species are equally abundant in each model.

by calculating the mean percentage abundance for each species present in either or both samples. The average sample was compared with all other samples and the revised indices of affinity were entered in the affinity matrix. The affinity matrix was again searched for the highest value and the sample pair giving this value was either placed in a new group, or, if one sample was an average sample for a group, the other sample was placed in the same group. If a group contained more than two samples, the average percentage abundance for each species in the group was found by adding together the species abundance for all samples in the group and dividing by the number of samples in the group. In other words, the average sample was an unweighted sample since all samples in the group were equally important in working out the average sample.

## ANALYSIS OF RESULTS

When text-figs. 2 and 3 are compared, it is found that the frequency distributions for the index of affinity are very similar to those for the modified coefficient of contingency. The chief difference lies in the curves for models where all the random samples are from the same artificial assemblage. The distribution in the case of the index of affinity is both lower in value and wider in range than is the distribution of the modified coefficient of contingency. When samples are taken from two artificial assemblages the frequency curves of the two measures of similarity are in general




TEXT-FIG. 3. Curves showing the frequency distribution of the modified coefficient of contingency ($D$) for sample pairs in artificial assemblages. The numbers at the peak of each curve indicate the number of species in the assemblage(s). Unless otherwise indicated, the species are equally abundant in each model.

identical. The differences which occur are due to the fact that a squared term occurs in the calculation of the modified coefficient of contingency, and this lowers the value of this coefficient when large numbers of 'species' are present in low abundance. Thus, when samples containing two species at 20% and five species at 4% are compared, the calculation for the coefficient of contingency involves comparing three species at 20% and fifteen at 4%. Thus the value of $\chi^2$ is high because $(X-EX)^2/EX$ is high when $EX$ is small, and $(X-EX)^2/EX$ is evaluated fifteen times for low values of $EX(EX=4\%)$ (where $X=$ observed value and $EX=$ the expected value). Although the curves are not the same for the two similarity indices under discussion, their differences may in general be predicted.

Having established that the index of affinity is a reliable indicator of similarity, let us look more closely at the distribution. Table 1 shows the range and peak of the frequency distribution curve for 95% of the sample pairs in each model. (The remaining 5% of the sample pairs have an index of affinity outside the range noted.)

For each model, the range is, approximately, 10. In the case of single assemblage models, the peak lies at about $(100-Y\times S)$ where $S$ is the number of species. $Y$ is a number whose value is dependent on the value of $S$. $Y$ is roughly equal to one, but

is more precisely a little greater than one when few 'species' are represented and slightly less than one when more than ten species are involved.

In the models where two types of assemblage are compared, the value of the peak of the frequency distribution curve is usually found to be the value which would be obtained if each sample in a sample pair was an accurate representation of the artificial assemblage. Thus, if one series of samples contains five species, each at 20%, and the second series of samples contains thirty species, each at $3\frac{1}{3}$%, the value of the index of affinity, if each sample is an exact replica of the model, would be $5 \times 3\frac{1}{3}$, i.e. $16\frac{2}{3}$. However, in the models for twenty *v.* thirty species and for two at 20% and five at 4%, the peak of the frequency distribution curves are found to be a little below the expected value—63 instead of 66 for twenty *v.* thirty species and 54 instead of 60 for two at 20% and five at 4%. Thus, if samples with large numbers of species are being compared, we may assume that similarity is implied when a rather lower value than expected is obtained.

TABLE 1. The value of the index of affinity for each model sample.

| Number of species in model | Lower | Peak | Upper | Range | Expected peak |
|---|---|---|---|---|---|
| 5 species | 88 | 94 | 97 | 9 | |
| 10 species | 83 | 89 | 94 | 11 | |
| 20 species | 80 | 84 | 88 | 8 | |
| 30 species | 74 | 81 | 86 | 12 | |
| 3 at 20%, 10 at 4% | 82 | 88 | 94 | 12 | |
| 30:5 | 12 | 17 | 21 | 9 | $16\frac{2}{3}$ |
| 20:5 | 20 | 26 | 29 | 9 | 25 |
| 10:5 | 44 | 50 | 56 | 12 | 50 |
| 20:10 | 44 | 50 | 56 | 12 | 50 |
| 30:10 | 27 | $33\frac{1}{3}$ | 39 | 12 | $33\frac{1}{3}$ |
| 30:20 | 58 | 63 | 68 | 10 | $66\frac{2}{3}$ |
| 2 at 20%, 5 at 4% | 48 | 54 | 59 | 11 | 60 |

Table 1 summarizes these results; note that the upper and lower $2\frac{1}{2}$% of the range of frequency distribution has been cut off, because those sample pairs are considered to be rather extreme.

If sample pairs from assemblages containing only five species have an index of affinity of 90 or more, we see from text-fig. 2 that this is well within the range of distribution expected for a five species example. In the case of the sample pairs (each containing about five species) taken by Buzas (1965) referred to earlier, the index of affinity for pairs of sample assemblages containing more than 100 individuals was usually 89 or greater, and we assume they were from the same assemblage. If we examine the two sample pairs giving values of less than 89, we find that one (Buzas's living, 10, 10′) was from a transitional zone where rapid changes may be expected, and the second (total assemblage, 104, 104′) contained only 115 individuals.

The value of the modified coefficient of contingency for each of Buzas's sample pairs is $> 95$ (except in living 10, 10′) which lies within the range expected for samples of five species.

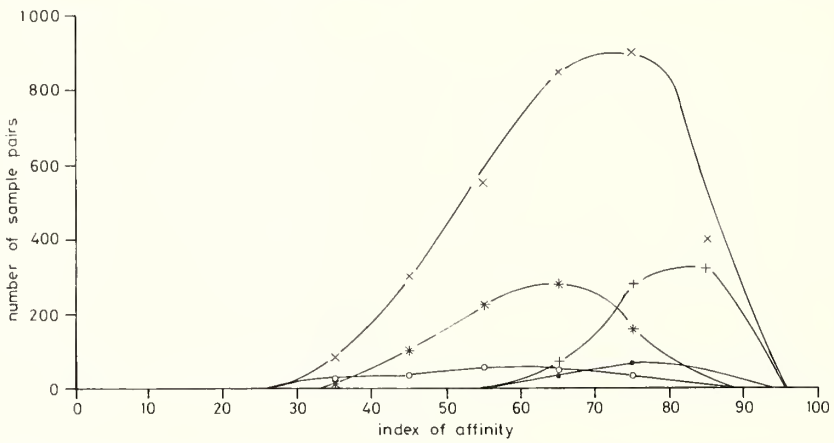
## HOLOCENE SEDIMENTS FROM THE WESTERN APPROACHES TO THE ENGLISH CHANNEL

A set of eighty surface and short-core samples of Holocene sediments containing planktonic Foraminifera were taken from the submarine canyon area of the Western Approaches to the English Channel. The water depth ranged from 100 to 2000 fathoms. From each sample the author counted about 250 individuals.

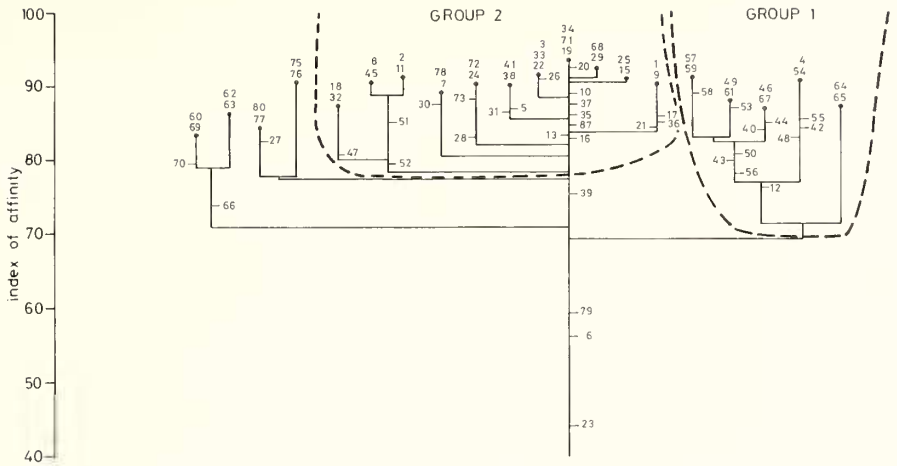
The assemblages were composed generally of about five species at 10–30% abundance and up to ten species at 5%, or, in many cases $< 1\%$ abundance. The range of frequency distributions is from 22 to 95 with a peak at about 70 (see text-fig. 4). If the samples followed the single assemblage model, they should produce a graph similar to the 20- or 10-species model, with a range about 80–95 with a peak of about 87 (see text-fig. 2). Since we are dealing with natural assemblages, lower values for the index of affinity may be accepted to indicate a similar origin for the samples.

In the present case, a range of 74–90 with a peak at 82 would indicate that the samples were from the same assemblage. The Holocene samples differ from this expected distribution because many sample pairs have an index of affinity lower than 75 which suggests that more than one assemblage is being sampled.

A grouping of the samples using the unweighted pair-group method described earlier is shown in text-fig. 5. Ideally, the groups should form above 75 if they contain samples from the same assemblage. Samples 64 and 65 were added to group 1 because they fell naturally near that group. Samples down to 39 could have been included in group 2. Within-group and between-group frequency distribution curves are shown for groups 1 and 2 (text-fig. 4). The within-group range for both groups is from 55 to 90, but the peak for group 2 is higher (about 82) than for group 1 (about 75) which indicates that group 1 is less homogeneous than group 2. The




TEXT-FIG. 4. Curves showing the frequency distribution of the index of affinity for planktonic Foraminifera from the Western Approaches. × = all samples; ● = group 1; + = group 2; ○ = ungrouped samples; * = group 1 compared with group 2.

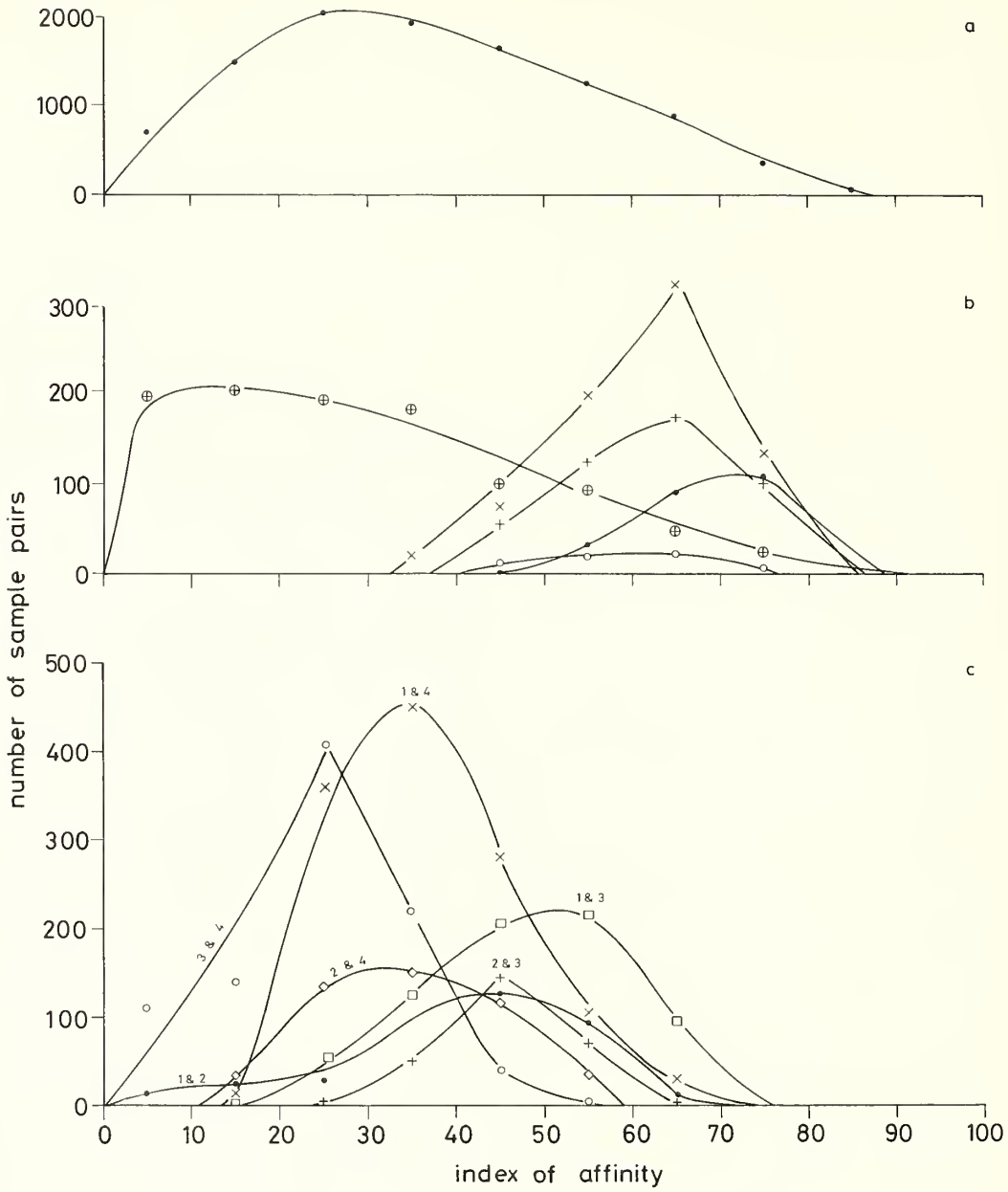



TEXT-FIG. 5. Grouping of Holocene samples of planktonic Foraminifera from the Western Approaches to the English Channel.

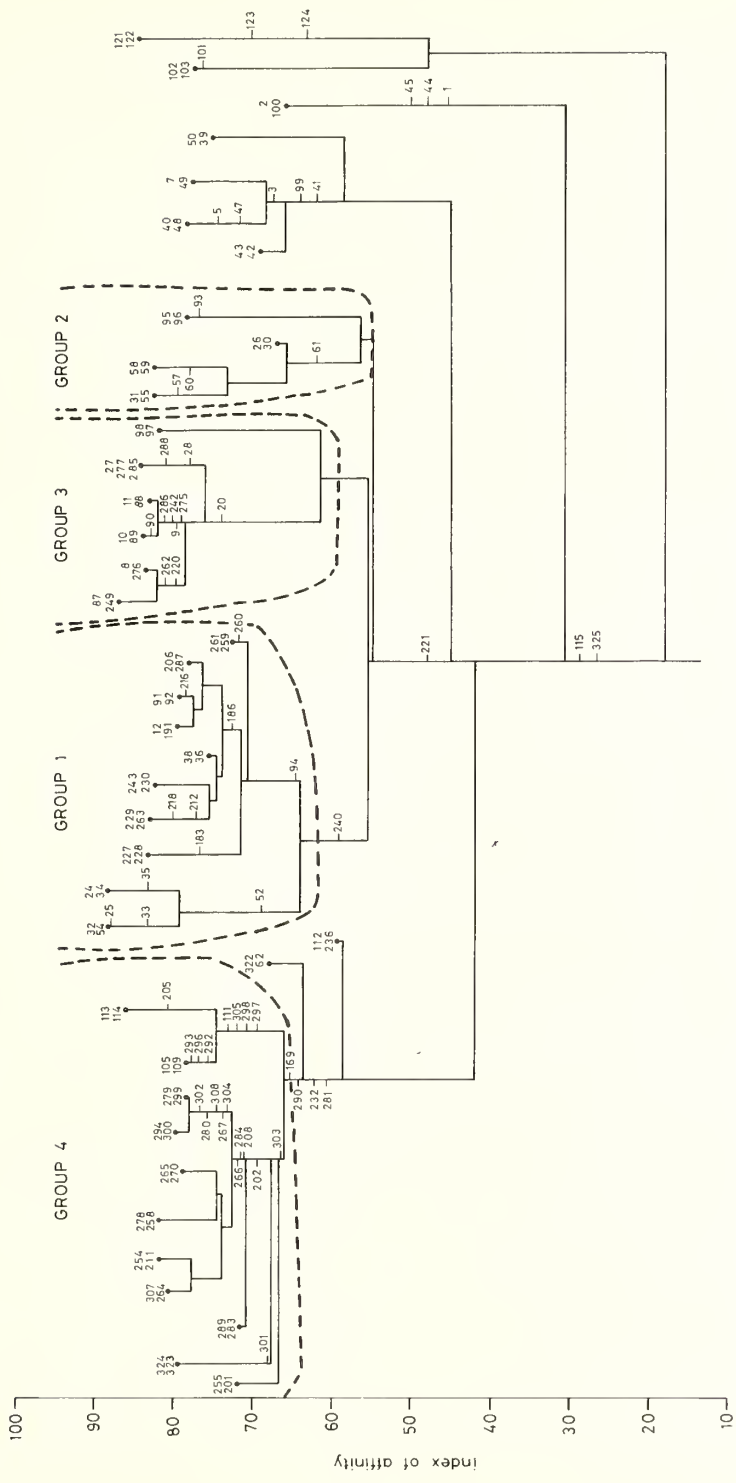
between-group curve shows a wider range (35–90) with a peak at 65 which suggests that the two groups are only moderately similar. The ungrouped samples have little in common with each other since the index of affinity for most sample pairs is less than 75.

## SAMPLES FROM THE CALIFORNIAN SEABOARD

Uchio (1960) analysed 154 samples of benthonic Foraminifera from the Californian seaboard, depths of the samples ranging from 0 to 600 fathoms. Up to sixty species were present in the dead (total) assemblage for each sample studied and at least 100 specimens were counted. Five species were predominant, others often occurring at $< 1\%$ abundance. As the assemblages lived on the sea floor, differences may well be found in neighbouring samples because of minor differences in the environment and because benthonic forms are usually patchily distributed. Thus, two samples which represent essentially the same environment may have two dominant species in common, but each may have an additional third dominant species not found in the other sample (see figs. 95, 96 in Murray 1973). Looking at the frequency distributions for the artificial assemblages, one might expect samples from the Californian seaboard to show a distribution similar to the 2 at 20%, 5 at 4% curve, or 20 species present out of 30 equally abundant species, i.e. a distribution ranging from 45 to 70 with a peak at about 55. In fact the frequency distribution for the index of affinity for all Uchio's samples ranges from 0 to 88 with a peak at 25 (text-fig. 6*a*). As one might expect, more than one assemblage was sampled. The author grouped the samples by the method indicated earlier, and four main groups were considered to be present (text-fig. 7). The curves of the frequency distribution for within- and between-group indices of affinity were drawn and are shown in text-fig. 6*b*, *c*. In each case the range for the within-group frequency distributions was from 30 to 90 with a peak $> 55$. The ungrouped samples gave a frequency distribution ranging from 0 to 80 with a peak at about 10.




TEXT-FIG. 6. Curves showing the frequency distribution of the index of affinity for benthonic Foraminifera (total assemblages) from the Californian seaboard (Uchio 1960). *a*, all samples compared with each other. *b*, within-group comparison of groups 1–4 and ungrouped samples. ○ = ungrouped samples; + = group 1; ○ = group 2; ● = group 3; × = group 4. *c*, groups 1–4 compared with each other.



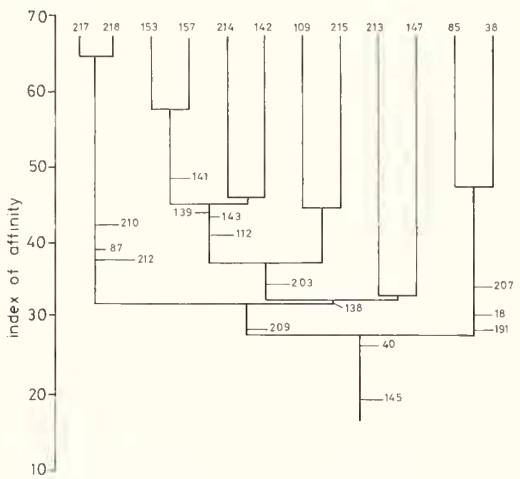

TEXT-FIG. 7. Grouping of samples of benthonic Foraminifera (total assemblages) from the Californian seaboard (Uchio 1960).

The between-group frequency distributions (text-fig. 6*c*) have a much lower peak —25 for groups 3 and 4, 55 (the highest between-group peak) for groups 1 and 3, which we considered to be most closely related in the clustering.

When these groups were compared with Uchio's depth zones, group 4 contained samples only from fauna 4 and indeed all the fauna 4 samples with the inclusion of sample 62 could have been grouped together. The fine-sand and coarse-sand distinctions were found and these two faunas were not linked. Uchio's faunas 2 and 3 were not revealed when the index of affinity was used for grouping. Group 2 corresponds to the deeper samples from the Loma Sea Valley and northern samples from 20 to 60 fathoms form a group that is not numbered. Group 3 contains most of the remaining samples from less than 100 fathoms and group 1 samples are in general from 100 to 200 fathoms.

## LIVING BENTHONIC FORAMINIFERA FROM THE GULF OF CALIFORNIA

Phleger (1964) described seventy-six samples of living benthonic Foraminifera from the Gulf of California. Of these only twenty-seven samples contained $> 100$ foraminiferids. Ten to thirty species were recognized in each sample, about four species being $> 10\%$ abundant. The composition of the sample assemblages is thus similar to those described by Uchio and if samples were drawn from the same benthonic assemblage the index of affinity should range again from about 45 to 70 with a peak about 55. In fact, when the frequency distribution of the index of affinity is calculated for the twenty-seven samples containing 100 specimens or more it is found that the range is 0–65 with a peak between 10 and 20. Only twenty-two sample pairs have an index of affinity $> 40$. This suggests that a wide variety of foraminiferal assemblages are present in the Gulf of California and Phleger did not get many samples from the same assemblage. Using Q-mode (sample/sample) factor analysis on all seventy-six samples, Streeter (1972) recognized nine assemblages and the characteristic species, together with their abundance are given for several of them. Groups obtained using the index of affinity were almost exactly the same as Streeter's groups, both in samples contained and, taking the mean species abundance for each group, in species content. However, only two sample pairs, 218, 217 and 153, 157, seem to represent two natural assemblages (see text-fig. 8) and it is doubtful whether further grouping of these samples adds to our understanding of the distribution of living benthonic Foraminifera in the Gulf of California.




TEXT-FIG. 8. Grouping of samples of living benthonic Foraminifera from the Gulf of California (Phleger 1964).

In recent years many numerical methods have been used to try to group

natural assemblages and in most cases the results support and amplify traditional techniques. The index of affinity would seem to be as effective as more complex methods and takes about a quarter of the time to compute as the complex methods. It may equally well be applied to small or large data sets. The range indicating that samples are from the same natural assemblage varies with the number of species present in quite a simple manner. It has been used here on modern and Holocene assemblages where the environment is known, but it should be possible to apply it to older fossil assemblages.

*Acknowledgements.* The author thanks Dr. P. J. Clifford and Dr. J. K. Ord for useful discussions on statistical problems and Professor J. W. Murray for much helpful criticism. Mr. D. Hamilton kindly made available the samples of Holocene sediment from the Western Approaches. The numerical work was done on the ICL 4-75 computer at Bristol University using the programming language AlgolW. A Computer Centre Library Procedure was used to generate the pseudo-random numbers.

## REFERENCES

BIRKS, H. J. B. 1974. Numerical zonations of Flandrian pollen data. *New Phytol.* **73,** 351–358.

BUZAS, M. A. 1965. The distribution and abundance of Foraminifera in Long Island Sound. *Smithson. misc. Collns*, **149,** 1–89.

CAIN, A. J. and HARRISON, G. A. 1958. An analysis of the taxonomist's judgement of affinity. *Proc. Zool. Soc. Lond.* **131,** 85–98.

CORMACK, R. M. 1971. A review of classification. *Jl R. statist. Soc.* **A 134,** 321–367.

CZEKANOWSKI, J. 1909. Zur Differential diagnose der Neandertalgruppe. *KorrespBl. dt. Ges. Anthrop.* **40,** 44–47.

GORDON, A. D. and BIRKS, H. J. B. 1972. Numerical methods in Quaternary palaeoecology. I. Zonation of pollen diagrams. *New Phytol.* **71,** 961–979.

—— —— 1974. Numerical methods in Quaternary palaeoecology. II. Comparison of pollen diagrams. Ibid. **73,** 221–249.

KENDALL, M. G. and STUART, A. 1967. *The advanced theory of statistics. Vol. 2. Inference and relationship.* London. 690 pp.

MURRAY, J. W. 1973. *Distribution and ecology of living benthic Foraminifera.* London. 274 pp.

PHLEGER, F. B. 1964. Patterns of living benthonic Foraminifera, Gulf of California. *Mem. Am. Assoc. Petrol. Geol.* **3,** 377–394.

SANDERS, H. L. 1960. Benthic studies in Buzzards Bay. III. The structure of the soft-bottom community. *Limnol. Oceanogr.* **5,** 138–153.

SNEATH, P. H. A. and SOKAL, R. R. 1973. *Numerical Taxonomy.* San Francisco. 573 pp.

SOKAL, R. R. and SNEATH, P. H. A. 1963. *Principles of Numerical Taxonomy.* San Francisco. 359 pp.

STREETER, S. S. 1972. Living benthonic foraminifera of the Gulf of California, a factor analysis of Phleger's (1964) data. *Micropaleontology*, **18,** 64–73.

UCHIO, T. 1960. Ecology of living benthonic Foraminifera from the San Diego, California area. *Spec. Publs Cushman Fdn*, **5,** 179–258.

M. J. ROGERS

Department of Geology
University of Bristol
Bristol, BS8 1TR

Typescript received 12 March 1975
Revised typescript received 3 October 1975